論文 / 著書情報
Article / Book Information

論題(和文)	音響特徴を用いた映像からのイベント検出の研究
Title(English)	Event Extraction from Video Data Harnessing Acoustic Features
著者(和文)	斉藤辰彦, 井上中順, 篠田浩一, 古井貞熙
Authors(English)	Tatsuhiko Saito, Nakamasa Inoue, Koichi Shinoda, SADAOKI FURUI
出典(和文)	日本音響学会2010年春季講演論文集, Vol. , No. , pp. 201-202
Citation(English)	, Vol. , No. , pp. 201-202
発行日 / Pub. date	2010, 3

音響特徴を用いた映像からのイベント検出の研究*

☆斉藤辰彦, △井上中順, 篠田浩一, 古井貞熙 (東工大)

1 はじめに

映像データベースの中からイベントを検索するためには, 映像がシーンごとにインデキシングされている必要がある. しかし, インデキシングの作業は人手で行われることが多く, そのコストは大きい. そのため, インデキシングを自動で行う技術が切望されている. 本研究では, 映像データベースの中から「人が歌っている」や「女性の顔のアップが映っている」といったイベントを検出するタスクに対して, 音響特徴を動画像特徴と組み合わせてインデキシングを行う手法を提案する.

2 映像からのイベント検出

映像検索を研究対象としたワークショップとしてTRECVID[1] がある. ここでは, その中の高次特徴 (High-level feature) 検出を対象とする. このタスクは映像中のショット (カメラの切り替わりが無い連続したフレームの集合) 単位ごとに特定の高次特徴が出現するか否かを判定することを目的とする. 高次特徴とは「船」や「椅子」などの物体, 「夜」や「町並み」などのシーン, あるいは「人が歌っている」などのイベントを指す. 高次特徴検出における従来研究では, 動画像特徴のみを用いたものがほとんどで, 音響特徴を用いた研究は少ない. 野球中継番組に対しては, 音響情報を用いてシーン識別精度を向上させる研究がある [2]. 本研究では, Singing や Person-playing-a-musical-instrument などの, 音響による効果が大きいと思われるイベントの検出に対して, 音響特徴を動画像特徴と組み合わせることによって検出精度の向上を目指す.

3 音響特徴を用いたイベント検出

3.1 音響特徴

まず, 各ショットから特徴量を抽出する. 本研究ではMFCC を用いる. そして, 検出対象の高次特徴それぞれに対して音響モデルを作成する. 音響モデルは隠れマルコフモデル (HMM) を用い, HMM の各状態に対するシンボルの出力確率分布には混合ガウス分布を用いる. 各イベントに対する HMM と, 全ショットに対する HMM (UBM; Universal Background Model) をそれぞれ作成し, 以下の式で表される対数尤度差 L_{A} に基づいてイベントの検出を行う.

$$L_{\mathrm{A}} = l_{\mathrm{HLF}} - l_{\mathrm{UBM}} \tag{1}$$

ここで, l_{HLF} はイベントに対する HMM から求めた対数尤度, l_{UBM} は UBM から求めた対数尤度である.

3.2 動画像特徴

動画像の特徴量は SIFT 特徴 [3] を用いる. 各ショットの全フレームから SIFT 特徴を抽出し, それぞれのショットに対して SIFT 混合ガウス分布 [4] を求める. ただし, 時間が短いショットではパラメータの推定に十分な量の SIFT 特徴が得られない可能性があるため, 映像全体の SIFT 混合ガウス分布を求め, それを基にした最大事後確率 (Maximum A Posteriori; MAP) 適応を各混合成分の平均ベクトルに対してのみ行うことで, ショットごとの SIFT 混合ガウス分布を求める. 次に, SIFT 混合ガウス分布間の距離を計算し, それを用いたカーネル SVM で学習を行う. 検出時には, SVM の出力を用いた事後確率推定により, 各ショットに対するイベントの出現確率を求め, 対数尤度差を得る.

3.3 音響特徴と動画像特徴の融合

音響特徴による対数尤度差を L_{A}, 動画像特徴を用いて求めた対数尤度差を L_{V} としたとき, 合成対数尤度差 L は以下のように求める.

$$L = w_{\mathrm{A}} L_{\mathrm{A}} + w_{\mathrm{V}} L_{\mathrm{V}} \tag{2}$$

ただし, w_{A}, w_{V} は $w_{\mathrm{A}} + w_{\mathrm{V}} = 1$ を満たす重み係数であり, w_{A} を 0 から 1 まで 0.1 刻みで変化させ, 高次特徴ごとの最適重みを事後的に求める.

4 評価実験

4.1 実験条件

評価実験では, TRECVID2009 の Development set として提供されている 100 時間の映像データを学習用・テスト用に二分割して用いた. このデータはオランダのドキュメンタリー番組や教育番組が主となっている. ショットの総数は学習用が 18,120, テスト用が 18,142 であり, 映像のショット境界と各ショットに対する高次特徴の出現の有無がラベルとして与えられている. なお、1 つのショットに対して複数のイベントが出現することを許している.

音響分析は 16kHz サンプリング, 25ms ハミング窓, 10ms フレームシフトで行い, 特徴量は 12 次元 MFCC, ΔMFCC, ΔΔMFCC, Δlog-power, ΔΔlog-power の 38 次元を用いた. また, 音響モデルは, 状態数 2, 混合数 512 のエルゴディック型 HMM を用いた. これらの

*Event Extraction from Video Data Harnessing Acoustic Features, by Tatsuhiko SAITO, Nakamasa INOUE, Koichi SHINODA, and Sadaoki FURUI (Tokyo Institute of Technology)

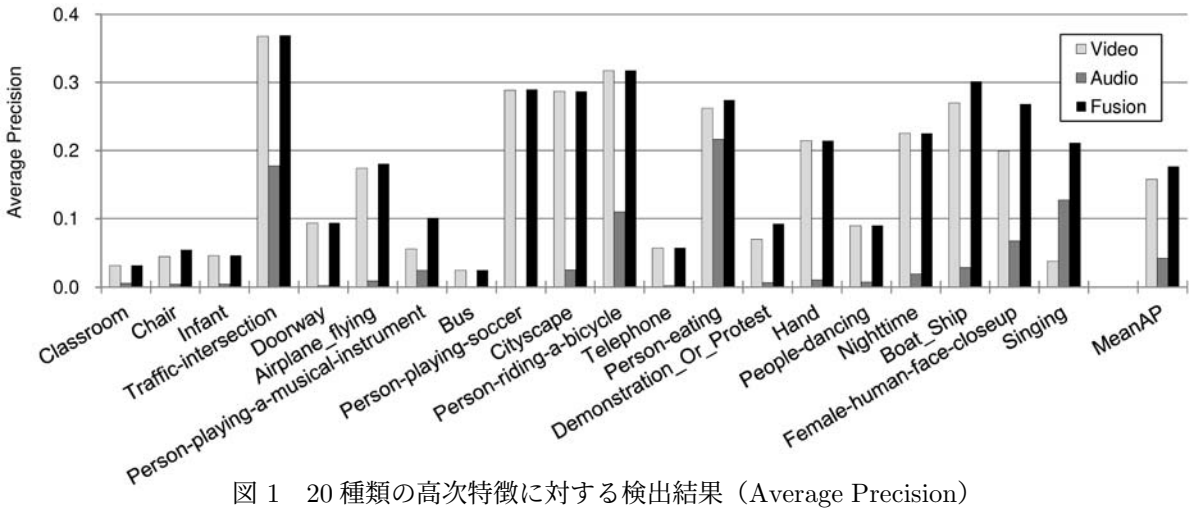

図 1　20 種類の高次特徴に対する検出結果（Average Precision）

状態数, 混合数は小規模データを用いた予備実験により決定した.

検出結果の評価尺度として, 各高次特徴に対する Average Precision (AP) と全高次特徴の AP を平均した MeanAP を用いる. AP は順位つき検出結果を評価する際によく用いられる評価尺度で, Precision-Recall 曲線と座標軸に囲まれた部分の面積に相当する. その計算式は以下のように表される.

$$\mathrm{AP} = \frac{1}{R}\sum_{r=1}^{N} \mathrm{Pr}(r)\mathrm{Rel}(r) \tag{3}$$

ここで, R は正解の総数, N は検索結果の総数, $\mathrm{Pr}(r)$ は第 r 位までの検出結果における Precision, $\mathrm{Rel}(r)$ は第 r 位の検出結果が正解であった場合に 1, そうでない時に 0 をとる関数である. $N = 2000$ として評価を行った.

4.2　実験結果

20 種類の高次特徴に対する AP とその平均の MeanAP を図 1 に示す. Video は動画像特徴のみを用いた結果, Audio は音響特徴のみを用いた結果, Fusion はこれらを組み合わせた場合の結果である. MeanAP は, 動画像特徴のみの場合で 0.158 となり, 音響特徴を合わせることで 0.176 まで向上した. Singing では音響特徴のみを用いた場合でも, 動画像特徴を用いた場合の倍以上の検出精度が得られている. 歌っているかどうかは動画像だけでは判断しにくく, 歌声を学習したと考えられる. Female-human-face-closeup では, 音響特徴を動画像特徴と組み合わせることで AP が顕著に向上した. 動画像特徴のみでは男性の顔と女性の顔の区別が難しいが, 音響特徴を組み合わせることで, 声により男性と女性の識別が可能となり, 結果として男性の顔のショットを排除することができたためと考えられる. また, Boat_Ship, Person-playing-a-musical-instrument では音響特徴のみでは動画像特徴のみを用いた場合に比べて低い結果が出ているものの, 両者を組み合わせることによって検出精度が大きく向上し, 音響特徴の有意性が示された. 一方で, Chair や Doorway など, 音響には直接的に関係しない高次特徴については, 音響特徴の効果は得られなかった.

5　おわりに

本研究では, 動画像特徴に加えて音響特徴を組み合わせるイベント検出手法を提案した. 高次特徴ごとに音響特徴を HMM でモデル化し, それによる対数尤度比と動画像特徴による対数尤度比を組み合わせることで, 検出精度がよくなることが示された. TRECVID2009 では, 同手法を用いた結果, Mean Inferred Average Precision が 0.168 となり, 参加 40 チーム中第 4 位の検出精度となった. 特に Singing と People-dancing では検出精度が第 1 位となり, 音響特徴の有意性が示された. 今後の課題としては MFCC 以外の特徴量, HMM 以外のモデル, また, 音声認識結果のテキスト特徴を用いることなどが挙げられる.

参考文献

[1] A. F. Smeaton *et al.*, “Evaluation campaigns and TRECVid”, MIR ’06: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval, pp. 321-330, 2006.

[2] 宮崎太郎 他, “野球中継番組を対象とした音響情報を用いたシーン認識”, 日本音響学会 2006 年春季講演論文集, pp. 19-20, 2006.

[3] D. G. Lowe, “Distinctive Image Features from Scale-Invariant Keypoints”, In Proc. International Journal of Computer Vision, vol. 20, pp. 91-110, 2004.

[4] 井上中順 他, “SIFT 混合ガウス分布と音響特徴を用いた映像からの高次特徴検出”, 電子情報通信学会技術研究報告, vol. 109, no. 306, pp. 97-102, 2009.